Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Multi-Objective Adaptive Feature Fusion Tracking Method for Autonomous Driving in Complex Occlusion Scenarios

Yongjie Zhang

yjzhang@ldy.edu.rs

Guangxi Vocational Normal University

Yanfeng Wu

Guangxi Vocational Normal University

Zi Gao

Guangxi Vocational Normal University

Zeyue Li

Guangxi Vocational Normal University

Yumei Zhou

Guangxi Vocational Normal University

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11739367/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Multi-Objective Adaptive Feature Fusion Tracking Method for Autonomous Driving in Complex Occlusion Scenarios

Yongjie Zhang[1,*], Yanfeng Wu[1], Zi Gao[1], Zeyue Li[1], Yumei Zhou[1]

[1]Guangxi Vocational Normal University, Xixiangtang, Nanning, 530007, Guangxi, China.

*Corresponding author(s). E-mail(s): yjzhang@ldy.edu.rs;
Contributing authors: yanfengwu2026@outlook.com;
zigao2026@outlook.com; zeyueli2026@outlook.com;
ymzhou2026@outlook.com;

Abstract

In complex occlusion scenarios, multi-object tracking for autonomous driving faces challenges such as the degradation of appearance features, failed data association and trajectory breaks. Most existing mainstream methods employ static feature fusion weights for data association, relying on a globally uniform appearance update rate and trajectory loss tolerance time. They fail to adaptively perceive the real-time occlusion status of objects, resulting in frequent template contamination and identity switching during prolonged occlusion. This paper proposes an occlusion-aware adaptive multi-feature fusion tracking framework. Firstly, an occlusion-aware adaptive multi-feature association mechanism is introduced, utilising a continuous occlusion score to drive a dynamic multi-feature weighting mechanism, thereby enabling real-time adaptive adjustment of the fusion weights comprising appearance, motion and IoU. Furthermore, an occlusion-driven template update strategy is designed, which adjusts the update rate based on the occlusion score and enforces a freeze during severe occlusion to prevent feature drift and template contamination. Finally, a joint hierarchical trajectory management mechanism is established, with different loss tolerance times designed according to occlusion severity and trajectory confidence, thereby effectively recovering high-quality trajectories lost due to occlusion.As evidenced by the comprehensive experimental results on the KITTI and nuScenes datasets, the metrics of our method—including MOTA and IDF1—are significantly better than those of several state-of-the-art mainstream improved algorithms. On the KITTI dataset, MOTA reached 69.8 per cent, representing an improvement of

3.4 per cent over EKF-BoT-SORT, the best-performing method among the comparators; on the nuScenes dataset, MOTA reached 64.3 per cent, representing an improvement of 3.8 per cent over EKF-BoT-SORT, the best-performing method among the comparators. Research has shown that this dynamic occlusion-aware framework can effectively improve the tracking continuity and identity preservation of vehicle and pedestrian targets under complex occlusion conditions, providing a high-precision, real-time and reliable solution for perception in autonomous driving.

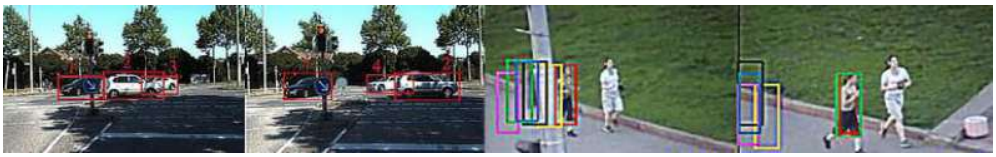
Keywords: multi-target tracking, autonomous driving, occlusion detection, adaptive feature fusion, hierarchical trajectory management

1 Introduction

As autonomous driving technology moves towards large-scale commercial deployment, the robustness of visual perception systems has become a core factor in ensuring the safe operation of intelligent vehicles. Complex backgrounds and occlusions pose significant challenges not only in the field of in-vehicle perspectives, but also in areas such as single-object tracking in satellite video. Kumar[1] et al. provide a detailed discussion of the impact of adverse weather conditions on bird's-eye-view perception; starting with the patterns of performance degradation in visual sensors under different meteorological conditions and the robustness of their architecture, they conduct a systematic analysis of how visual sensors operate in adverse weather conditions.Traffic footage captured by in-vehicle cameras in real-world urban environments is frequently subject to sudden changes in lighting, interference caused by adverse weather conditions such as rain, snow and fog, and complex background noise from multiple perspectives; this places higher demands on the robustness of perception systems. Lei[2] et al. proposed the EdgeCaseDNet detection architecture to address edge cases such as fog, rain, snow and night-time conditions, utilising feature enhancement to improve object detection performance under adverse weather conditions. Multi-object tracking (MOT) is a key component of the perception stage in autonomous driving systems; its function is to identify vehicles or pedestrians as targets and ensure they are tracked. Due to the frequent interactions between dynamic objects in urban scenarios, severe mutual occlusion between objects has become one of the primary challenges facing current MOT algorithms. Liu[38]et al. argue that appearance changes and trajectory breaks caused by occlusion constitute the primary obstacles to MOT, and propose the StrongerCentre framework, which combines occlusion awareness with cascaded matching to enhance the robustness of tracking. Occlusion not only causes significant changes to the target's appearance features but also leads to breaks in trajectory association, thereby affecting subsequent task planning[3]. Consequently, achieving high-precision, highly robust continuous multi-object tracking under complex, continuous and multi-level occlusion conditions remains a technical challenge currently faced by both academia and industry.



(a)



(b)



(c)

Fig. 1 Three typical complex occlusion challenges faced in autonomous multi-object tracking: (a) mutual occlusion between densely packed objects; (b) occlusion and appearance defects in static environments; (c) low visibility and feature degradation caused by extreme weather conditions.

Existing tracking-by-detection methods, such as DeepSORT and ByteTrack, have achieved good tracking performance in scenes with no occlusions or only minor occlusions. DeepSORT employs Kalman filtering and the Hungarian algorithm for data association, utilising a cost matrix with fixed weights for appearance and motion features for matching[5]; ByteTrack adopts a hierarchical association approach based on detection scores to improve trajectory recall in occlusion scenarios[32]; researchers have proposed a multimodal visual tracking framework that utilises complementary sensor information to enhance the system's discriminative capability in complex backgrounds[6]; To address the challenges of target appearance similarity and frequent entry and exit points, frameworks such as PoseTrack combine 3D pose estimation with temporal visual cues to strengthen cross-frame correlation; InteBOMB incorporates a background enhancement strategy and an online correction mechanism utilising long-short-term memory to maintain dynamic adaptation to target features[7]. However, these methods have key limitations in terms of occlusion awareness: firstly, the feature fusion weights are set statically and cannot be adapted according to the degree of occlusion; when occlusion occurs, appearance features degrade, yet the appearance metrics under fixed weights remain dominant, leading to erroneous associations[9][10]; secondly, the appearance template update strategy fails to effectively suppress the introduction of contaminated features during occlusion, resulting in the accumulation of feature drift[12]; thirdly, trajectory management employs a uniform loss tolerance threshold, overlooking differences in occlusion severity and confidence levels among individual trajectories; high-confidence trajectories are terminated prematurely

following severe occlusion, whilst low-confidence false-alarm trajectories persist for extended periods[11]. Furthermore, existing methods have blind spots across three dimensions—the degree of occlusion detection, feature fusion strategies and the level of trajectory management—and lack mechanisms for the quantitative detection of occlusion continuity and adaptive adjustment[13].

To address the aforementioned issues, this paper proposes an occlusion-aware adaptive multi-feature fusion tracking framework. Based on a three-dimensional evaluation space comprising "occlusion-aware association → occlusion-driven updates → joint hierarchical management", it systematically resolves the cascading problems of appearance feature degradation, data association failure and trajectory discontinuity in complex occlusion scenarios. The main innovative contributions of this paper are as follows:

(1) Firstly, we propose an occlusion-aware adaptive multi-feature correlation mechanism. This mechanism utilises a continuous occlusion score to drive dynamic multi-feature weighting, thereby adaptively adjusting the fusion weights comprising appearance, motion and IoU. When an object is occluded, the continuous occlusion score drives a softmax-normalised weight allocation, which adaptively downweights unreliable appearance features whilst emphasising more robust spatial position correlations, thereby achieving a precise alignment between the degree of occlusion and feature confidence.

(2) Secondly, we designed an appearance template update strategy driven by occlusion. This strategy dynamically scales the update rate based on the occlusion score, and forces a freeze during severe occlusion to suppress feature drift and template contamination. At the same time, we introduced a template quality assessment mechanism that triggers updates only when the detection confidence is sufficiently high; when the target becomes clearly visible again, the normal update rate is resumed to rapidly incorporate new appearance information, thereby effectively curbing the accumulation of feature drift during occlusion.

(3) Finally, a joint hierarchical trajectory management mechanism is established. This mechanism combines occlusion levels with trajectory confidence to design differentiated loss tolerance times, thereby classifying trajectories into six tiers. Trajectories with high confidence and severe occlusion are assigned the longest tolerance time to effectively recover high-quality trajectories lost due to occlusion; those with low confidence and no occlusion are assigned the shortest tolerance time to swiftly eliminate false alarms. This mechanism moves away from a one-size-fits-all trajectory termination strategy, significantly improving trajectory recovery rates in scenarios of severe occlusion whilst maintaining a low false alarm rate.

2 Materials and Methods

This section provides a systematic review of research into multi-object tracking for autonomous driving from three perspectives: an occlusion-aware adaptive multi-feature association mechanism, an occlusion-driven adaptive update strategy for appearance templates, and a trajectory management framework that jointly classifies

trajectories based on occlusion levels and trajectory confidence. It analyses the shortcomings of existing methods and presents the corresponding solutions proposed in this paper.

2.1 Occlusion-Aware Adaptive Multi-Feature Correlation Mechanism

Data association is a crucial step in multi-object tracking, the aim of which is to establish a reliable match between detection results and existing trajectories.The improved version of DeepSORT[34] employs a cascaded matching strategy, using a weighted sum of the cosine distance of appearance features and the Mahalanobis distance as the association cost, and solves for the cost-minimising match using the Hungarian algorithm.However, the fusion weights in this method (appearance weight 0.9, motion weight 0.1) are empirically fixed values; when the target is occluded, leading to the degradation of appearance features, unreliable appearance metrics still dominate, which can easily lead to misalignment errors . An improved version of ByteTrack[35] proposes a tiered association strategy based on detection scores, categorising detection boxes into two tiers—high and low—according to their scores. It prioritises matching high-score detection boxes before matching the remaining trajectories with low-score detection boxes.Although this strategy has, to some extent, improved recall in occlusion scenarios, it is essentially a matching strategy based on detection confidence rather than occlusion awareness, and lacks a quantitative assessment of the degree of occlusion.

In recent years, a number of research studies have emerged focusing on improvements to association in occlusion scenarios . Liu[4] et al. proposed the AMTrack algorithm, which quantifies the degree of target occlusion by constructing an Occlusion Index (ONI) and designs an Anti-Occlusion Multi-Level Association (AMLA) mechanism to perform hierarchical association of targets based on the degree of occlusion . However, its ONI is calculated solely on the basis of the overlapping regions of detection boxes, making it difficult to reflect the duration of occlusion and the extent of feature degradation. Huang[14] et al. introduced an occlusion-aware affinity calculation module within a camera-LiDAR fusion framework, utilising occlusion information in 3D space to enhance the robustness of the correlation; however, the fusion weights in this method are still set statically . Wang[10] et al. proposed the OATrack framework, which uses an occlusion-aware module to estimate the occlusion rate and adaptively adjusts the innovation gain and correlation coefficient weights of the Kalman filter according to the degree of occlusion . This method achieved satisfactory results on the MOT17 dataset; however, its occlusion-aware module only outputs a binary occlusion status, lacking a fine-grained quantification of occlusion continuity. Tan[15] et al. proposed a 3D multi-object tracking method based on parallel feature extraction and data association, which processes 2D appearance features and 3D spatial information separately; however, the fusion weights between appearance and motion features remain constant throughout the association process. Sun[16] et al. proposed the KDPBTracker, which utilises the design of virtual nodes and predicted confidence scores to improve the robustness of the data association matrix; however, the confidence levels of their virtual nodes are determined using empirical thresholds.

Consequently, most current matching methods rely on static feature fusion weights or binary occlusion detection; they lack the ability to quantify the continuity of occlusion or to adaptively adjust weights. When appearance features degrade under occlusion, these methods are unable to promptly adjust their matching strategies to downplay unreliable appearance metrics and emphasise spatial position metrics; there is therefore an urgent need for an adaptive multi-feature matching mechanism driven by a continuous occlusion score.

2.2 Adaptive update strategy for occlusion-driven appearance templates

The update strategy for appearance templates directly determines the impact of the tracker on the quality of features in occlusion scenarios. DeepSORT employs an exponential moving average (EMA) update method with a fixed learning rate to update both the newly detected appearance features and historical templates for each frame. When an object is occluded, appearance features from the occluder or background are continuously incorporated into the template, causing the template to become progressively contaminated and resulting in cumulative feature drift. ByteTrack links the EMA update rate to the detection score; however, it does not take the effect of occlusion into account.

Maintaining the purity of the target template is essential for ensuring the stability of long-term tracking. The GOT-JEPA framework proposed by Chen[8] et al. utilises joint embedding prediction to achieve model adaptation, employing pseudo-supervision to enhance the model's feature extraction capabilities in complex backgrounds. For multimodal features, Liu[17] et al. proposed MPTrack, a prototype-based learning approach, which employs prototype evaluation and adaptive update algorithms to manage templates. In recent influential research, Sakthi[18] et al. proposed DynaTrack, a single-object tracking method that selects an appropriate "confidence-gated dynamic memory" to selectively store reliable tracking targets whilst suppressing corrupted templates; R. Li[19] et al. proposed the Hierarchical Adaptive Feature Update (HAFU) method; Zhu[20] et al. proposed the SRRT framework, thereby introducing a method for determining updates based on a locked state to adapt to appearance changes occurring during the tracking process.

It is worth noting that, although DynaTrack introduces concepts such as "gating" and "confidence" , mainstream methods still employ a globally uniform, fixed learning rate when setting the template update rate. This means that when a target is heavily occluded, existing confidence-based gating cannot respond immediately. If a fixed learning rate is continued to be used for feature learning, the model can easily incorporate background clutter, noise or non-target features into the template within a short period of time, resulting in severe "template contamination" and feature drift, which in turn leads to ID Switch that is difficult to reverse. The confidence-based gating relied upon by DynaTrack cannot accurately determine the extent of dynamic occlusion.

To mitigate the issue of template contamination, this paper proposes an occlusion-driven method for updating appearance templates. This method overcomes the shortcomings of traditional global fixed learning rates by using the current real-time

degree of occlusion of the target as the primary criterion to adaptively adjust the update rate of the appearance template. When the model determines that the target is clearly visible, it maintains normal learning to capture subtle appearance changes; when the model detects that the target is severely occluded, it reduces the update rate or even halts updates. This adaptive adjustment mechanism, which utilises occlusion feedback, effectively suppresses feature drift and ensures the purity of the appearance template.

2.3 A Trajectory Management Framework Based on Joint Stratification of Occlusion Levels and Trajectory Confidence

Trajectory management primarily involves making decisions regarding the initialisation, termination and recovery of trajectories, and determines the integrity of the tracking system. DeepSORT employs a uniform threshold for the maximum allowable loss duration (`max_age`); if a trajectory remains lost for more than `max_age` consecutive frames, it is terminated[5].This strategy presents the following issues: high-confidence, high-quality trajectories may lose a few frames of detection following severe occlusion, yet their motion state remains reliable and their historical features are rich; they should therefore be afforded a longer waiting period. Conversely, low-confidence false-alarm trajectories should be deleted immediately, even if they are lost for only an instant. A uniform threshold cannot distinguish between these two scenarios. ByteTrack[32] employs a tiered management system based on detection scores, but its decision to terminate a track still relies on a uniform threshold for the number of lost frames.

With regard to adaptive trajectory management, Gamerdinger[21] et al. conducted a systematic analysis of trajectory management strategies in collaborative autonomous driving environments, concluding that a single loss tolerance time is insufficient to address multi-object scenarios. Research by Min[22] et al. demonstrated that motion estimation plays a crucial role in SORT; they proposed a new algorithm called OR-SORT, which utilises integer linear programming to eliminate outliers, thereby adapting to camera motion and improving the accuracy of trajectory prediction. Jin[9] et al. proposed a 3D multi-object tracking method that enhances data association and improves the trajectory management mechanism, using dynamically adjusted trajectory confidence to modify the management strategy; however, their confidence evaluation was applied solely to the detection score and did not take into account the impact of occlusion levels. Sun[16] et al. proposed a module in KDPBTracker that determines whether to retain a trajectory based on confidence, dynamically managing trajectory continuity according to thresholds for time and reliability; however, its two-dimensional hierarchical strategy still treats detection confidence as the primary evaluation dimension, with the level of occlusion playing only a supplementary role. To address the challenges posed by low-frame-rate video, Zeng[23] et al. employed the SORT-LFR method, which utilises an accelerated Kalman filter and DG-AEMA feature updates. Furthermore, in the context of autonomous driving deployment, Li[24] et al. proposed an interactive trajectory prediction framework to address perception failures, utilising an object lifecycle module to handle different issues in a targeted manner.

Consequently, existing trajectory management methods employ a uniform threshold for the permissible duration of loss of track, without taking into account the differences in occlusion levels and confidence levels between individual trajectories. This results in high-confidence trajectories being terminated prematurely, whilst low-confidence false alarms remain in the system; there is therefore an urgent need to establish a differentiated, hierarchical trajectory management approach that combines occlusion levels with trajectory confidence levels.

3 Methods

This paper proposes an adaptive, occlusion-aware multi-feature fusion tracking method to address the issues of feature degradation and policy rigidity in complex occlusion scenarios for autonomous driving. The framework utilises continuous occlusion scores as the primary driving factor, establishing a progressive processing sequence comprising "occlusion-aware linkage → occlusion-driven updates → joint hierarchical control" . The first section of this paper presents the complete structural framework, followed by a detailed introduction to each key module.

3.1 Overview of the Overall Framework

Multi-target tracking essentially involves achieving a consistent estimate of a target's identity and state within a dynamic spatio-temporal sequence. Drawing inspiration from the aerospace sector's approach to achieving high-precision trajectory tracking through dynamic adaptation and time-varying uncertainty estimation[26], as well as strategies in multi-agent systems that utilise joint trajectory and resource optimisation to address complex dynamic network environments[27], this paper proposes a framework for occlusion-aware adaptive multi-feature fusion tracking based on cascaded perception and dynamic decision-making. This framework utilises the continuous occlusion score as the primary driving signal, establishing a processing workflow comprising three interconnected stages: "occlusion-aware association" , "occlusion-driven updating" and "joint hierarchical management". This achieves end-to-end, occlusion-adaptive closed-loop control over data association, template preservation and trajectory lifecycle management. The overall architecture of the occlusion-aware adaptive multi-feature fusion tracking framework is shown in Figure 2.

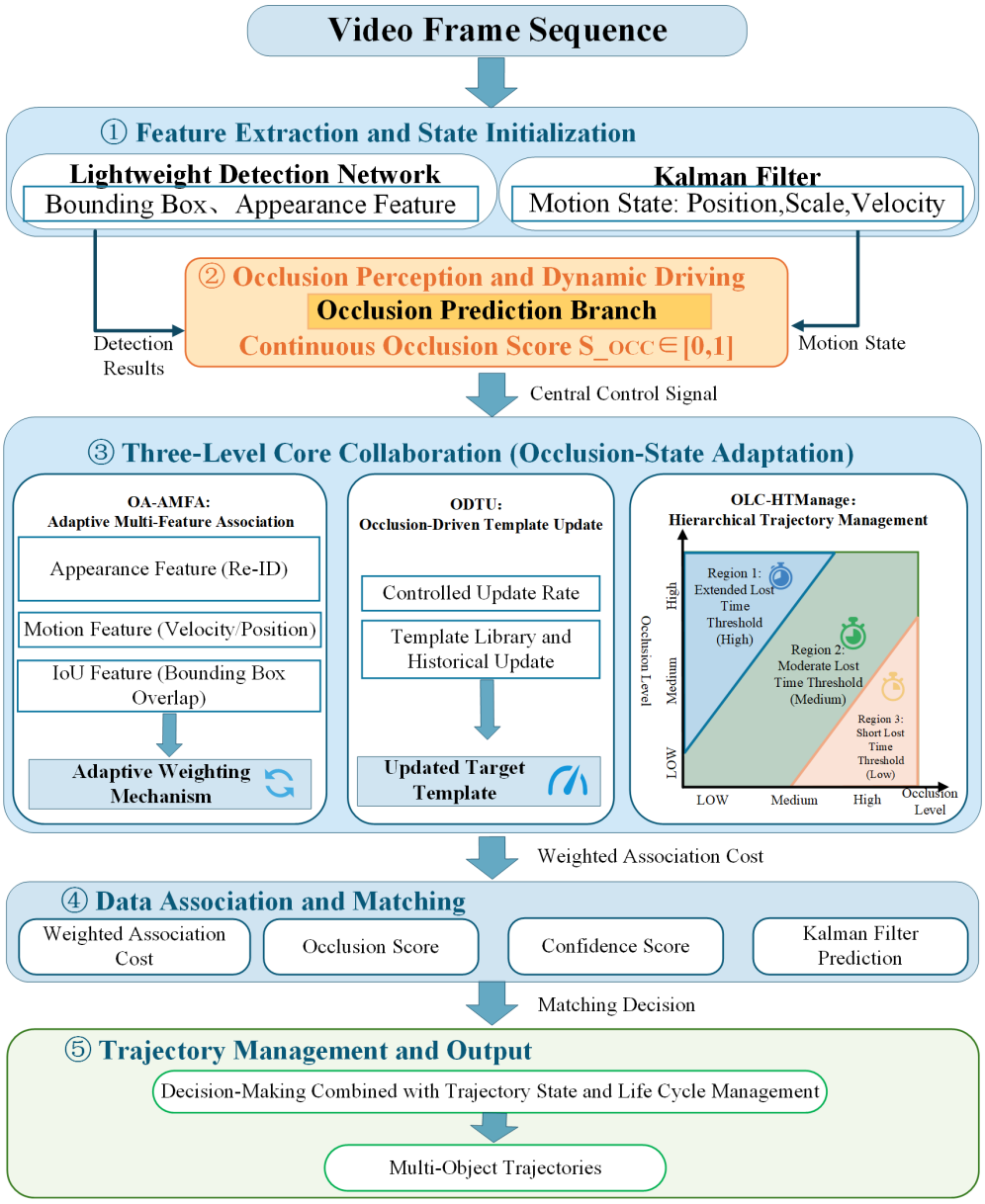


Fig. 2 Overall architecture diagram of the occlusion-aware adaptive multi-feature fusion tracking framework.

The overall workflow of the method described in this paper can be divided into four parts. The first stage involves feature extraction and state initialisation: a lightweight detection network is used to obtain the object's bounding box and appearance features, whilst a Kalman filter is employed to establish a motion state model (containing

information on position, size and velocity). The second stage involves occlusion detection and dynamic control: utilising a specially designed occlusion prediction branch, a continuous occlusion score $S_{\text{occ}}^{(\text{i})} \in [0,1]$ is calculated based on the target's detection confidence and the rate of change in its appearance features over time; this score is used as the "central control signal" for the respective adaptive modules in subsequent stages. The third stage involves three-level core coordination: $S_{\text{occ}}^{(\text{i})}$ is fed into the multi-feature association module, the template update module and the trajectory management module respectively, thereby achieving closed-loop control based on the occlusion state. The fourth stage involves trajectory output, which integrates the results of feature correlation and the decisions made by the trajectory management module to produce a robust and continuous tracking trajectory. For downstream tasks (such as autonomous driving trajectory prediction), the long-term, continuous nature of this trajectory output effectively mitigates cumulative errors caused by perception gaps[24].

3.2 Occlusion-Aware Adaptive Multi-Feature Correlation Mechanism

The OA-AMFA mechanism is one of the key innovations within the occlusion-aware adaptive multi-feature fusion tracking framework; when occlusion conditions are constantly changing, it adaptively adjusts the correlation cost weights assigned to each feature metric, thereby ensuring that they align with the reliability of each feature metric under the current occlusion conditions.

Suppose that frame t contains N active trajectories $\{T_i\}_{i=1}^{N}$ and M detection results $\{D_j\}_{j=1}^{M}$. For each trajectory T_i, we must first calculate its continuous occlusion score $s_i \in [0,1]$:

$$s_i = \alpha \cdot r_i + \beta \cdot \min\left(\frac{d_i}{d_{\max}}, 1\right) + \gamma \cdot (1 - c_i) \tag{1}$$

Here, r_i denotes the occlusion ratio, d_i denotes the number of consecutive frames of occlusion, $d_{\max}$ denotes the maximum number of effective occlusion frames, and c_i denotes the feature confidence; in this paper, the weighting coefficients α, β and γ are 0.5, 0.3 and 0.2 respectively.

Based on the occlusion score s_i, calculate the adaptive weights for the three types of feature metrics:

$$w_{\text{app}}^{(i)} = \frac{\exp(-\lambda_{\text{app}} \cdot s_i)}{\sum_{k\in\{\text{app, mot, iou}\}} \exp(-\lambda_k \cdot s_i)} \tag{2}$$

$$w_{\text{mot}}^{(i)} = \frac{\exp(\lambda_{\text{mot}} \cdot s_i)}{\sum_{k\in\{\text{app, mot, iou}\}} \exp(\lambda_k \cdot s_i)} \tag{3}$$

$$w_{\text{iou}}^{(i)} = \frac{\exp(\lambda_{\text{iou}} \cdot s_i)}{\sum_{k\in\{\text{app, mot, iou}\}} \exp(\lambda_k \cdot s_i)} \tag{4}$$

In this context, λ_{app}, λ_{mot} and λ_{iou} are scaling factors used to control the occlusion sensitivity of each metric (set at 2.0, 1.5 and 1.0 respectively in this paper). When s_i is very close to 0, w_{app} is at its maximum, and the correlation is dominated by appearance features; when s_i is very close to 1 (severe occlusion), w_{app} approaches 0,

and the correlation is dominated by w_{mot} and w_{iou}, with matching achieved through motion prediction and spatial overlap.

The element C_{ij} of the association cost matrix $C \in \mathbb{R}^{N \times M}$ is defined as:

$$C_{ij} = w_{\text{app}}^{(i)} \cdot d_{\text{app}}(T_i, D_j) + w_{\text{mot}}^{(i)} \cdot d_{\text{mot}}(T_i, D_j) + w_{\text{iou}}^{(i)} \cdot d_{\text{iou}}(T_i, D_j) \tag{5}$$

Here, d_{app} denotes the cosine distance based on visual features, d_{mot} denotes the Mahalanobis distance, and d_{iou} denotes the IoU distance. Finally, the Hungarian algorithm is employed to find the optimal matching for the cost matrix C. The workflow of the OA-AMFA mechanism is illustrated in Figure 3 below.

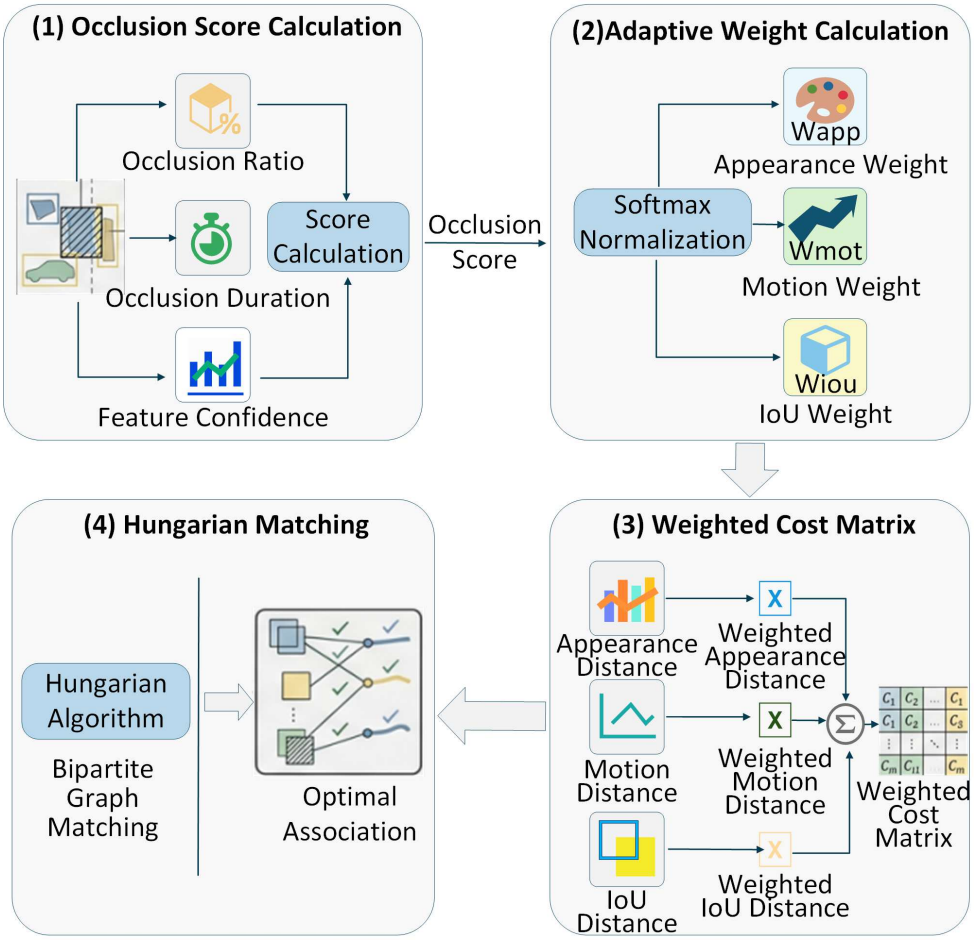


Fig. 3 OA-AMFA Mechanism Workflow.

3.3 Occlusion-Driven Appearance Template Update Strategy

Appearance templates are fundamental to target identification. Recent research has shown that trackers suffer from significant feature drift when processing deforming targets , they cannot guarantee that the template remains unaltered in the presence of disturbances[17]. However, most mainstream tracking algorithms[18] currently still use a globally fixed update rate, β_{base}, to fuse historical and current features. This crude update method readily introduces background noise into the template when the target is occluded, thereby causing template contamination. The workflow for the occlusion-driven appearance template update strategy is illustrated in Figure 4 below.

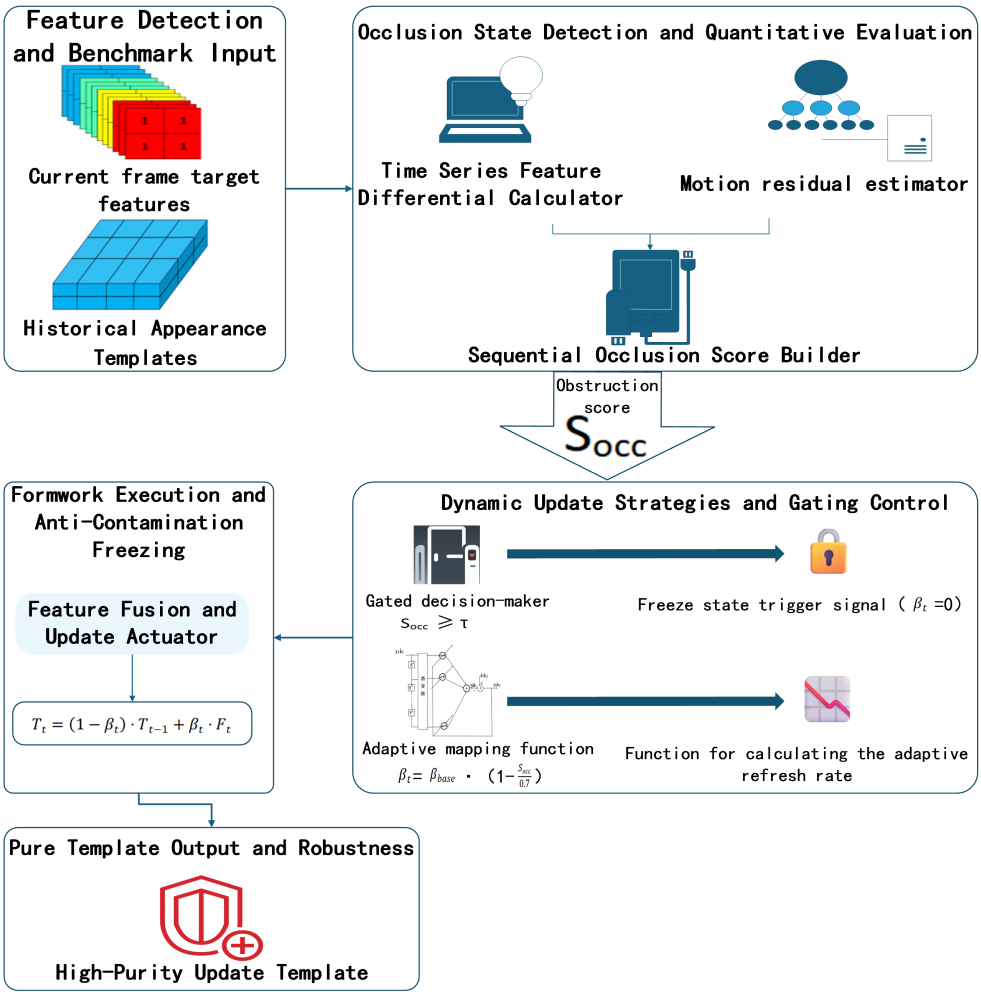


Fig. 4 SubsectionFlowchart of the occlusion-driven appearance template update strategy.

This strategy comprises five main functional domains: feature perception and benchmark input; occlusion state perception and quantification assessment; dynamic strategy updating and gating control; template execution and anti-contamination freezing; and clean template output. The core idea is to utilise dynamic occlusion scores to trigger a dual-path gating decision mechanism, thereby effectively addressing the issue of template contamination.

By incorporating a two-stage joint evaluation mechanism into "occlusion state detection and quantitative evaluation", the extent to which a target is occluded is estimated.Firstly, the temporal feature difference calculator computes the rate of decline in the cosine similarity of appearance features between two adjacent frames to assess appearance degradation; the motion residual estimator, meanwhile, analyses the Mahalanobis distance residual between the predicted position (derived from the Kalman filter) and the actual detected position to determine whether the target has undergone large-scale positional changes. These two metrics are then fed into the continuous occlusion score builder, where they are fused via a non-linear mapping to form a comprehensive dynamic occlusion score $S_{\text{occ}} \in [0, 1]$ .This score serves as a "control signal" for each module, providing real-time feedback on the severity of the target's occlusion.

The data stream then enters the "Dynamic Update Strategy and Gating Control" domain.To prevent background noise from being erroneously incorporated into the template under conditions of severe occlusion, gating is implemented by setting a hard trigger threshold $\tau = 0.7$ for severe occlusion: when $S_{\text{occ}} \geq \tau$, the decision maker directly triggers a freeze signal, forcibly suppressing updates; when $S_{\text{occ}} < \tau$, the adaptive update rate β_t is calculated based on the current S_{occ}. In the "Feature Awareness and Reference Input" domain, the system first obtains the target "current frame target feature" F_t and the "historical appearance template" T_{t-1}. The adaptive update rate β_t is expressed by the formula:

$$\beta_t = \begin{cases} \beta_{\text{base}} \cdot \sin\left(\frac{\pi}{2} \cdot \left(1 - \frac{S_{\text{occ}}(t)}{\tau}\right)\right), & S_{\text{occ}}(t) < \tau \\ 0, & S_{\text{occ}}(t) \geq \tau \end{cases} \tag{6}$$

Here, β_t is the adaptive update rate for frame t, with values ranging from $[0, \beta_{\text{base}}]$. β_{base} serves as the base update rate, i.e. the maximum learning rate, whilst $S_{\text{occ}}(t)$ represents the continuous occlusion score for frame t, and τ is the severe occlusion gating threshold. This strategy causes the template update rate to exhibit a smooth, non-linear decay as the degree of occlusion increases, thereby effectively ensuring the purity of the template.

Under the combined influence of the two control signals mentioned above, the process enters the "template execution and anti-contamination freeze" domain. The feature fusion and update actuator updates the template according to the following exponential moving average (EMA) formula:

$$T_t = (1 - \beta_t) \cdot T_{t-1} + \beta_t \cdot F_t \tag{7}$$

Here, T_t denotes the high-purity appearance template feature vector after the update in frame t, T_{t-1} denotes the historical appearance template feature vector in frame $t-1$, and F_t denotes the appearance feature vector of the current target in frame t.

Freezing template updates when occlusion occurs helps to maintain template purity and, to a certain extent, reduces the long-term drift and "ID Switch" issues caused by a fixed update rate. When an actuator receives a "freeze state trigger signal ($\beta_t = 0$)" from an upstream component, the update operation is immediately locked. Consequently, the features F_t of the current frame are not written to the historical template, thereby effectively suppressing the inclusion of noise features.

3.4 A Joint Hierarchical Trajectory Management Framework

Trajectory lifecycle management has become the final line of defence for maintaining long-term consistency in output. Traditional approaches[21] typically treat a single, globally uniform static parameter, Tbase , as the tolerance time for trajectory loss; this fails to address the wide variety of dynamic occlusion events encountered in autonomous driving. As the complex and ever-changing nature of intelligent transport environments becomes increasingly apparent, global planning should be combined with local dynamic execution to adapt to environmental changes[30].

In response to the shortcomings of the crude trajectory management approach outlined above, this paper proposes a hierarchical trajectory management framework based on multi-source state awareness and two-dimensional joint mapping. This framework is structured around four components: the multi-source state awareness domain, the key feature and metric quantification domain, the two-dimensional joint decision-making and conflict resolution domain, and the dynamic lifecycle execution domain. Unlike traditional methods that employ a single, globally uniform threshold, this framework assigns different loss tolerance times to different data streams based on the real-time dynamic state of the trajectory; the specific structure is illustrated in Figure 5.

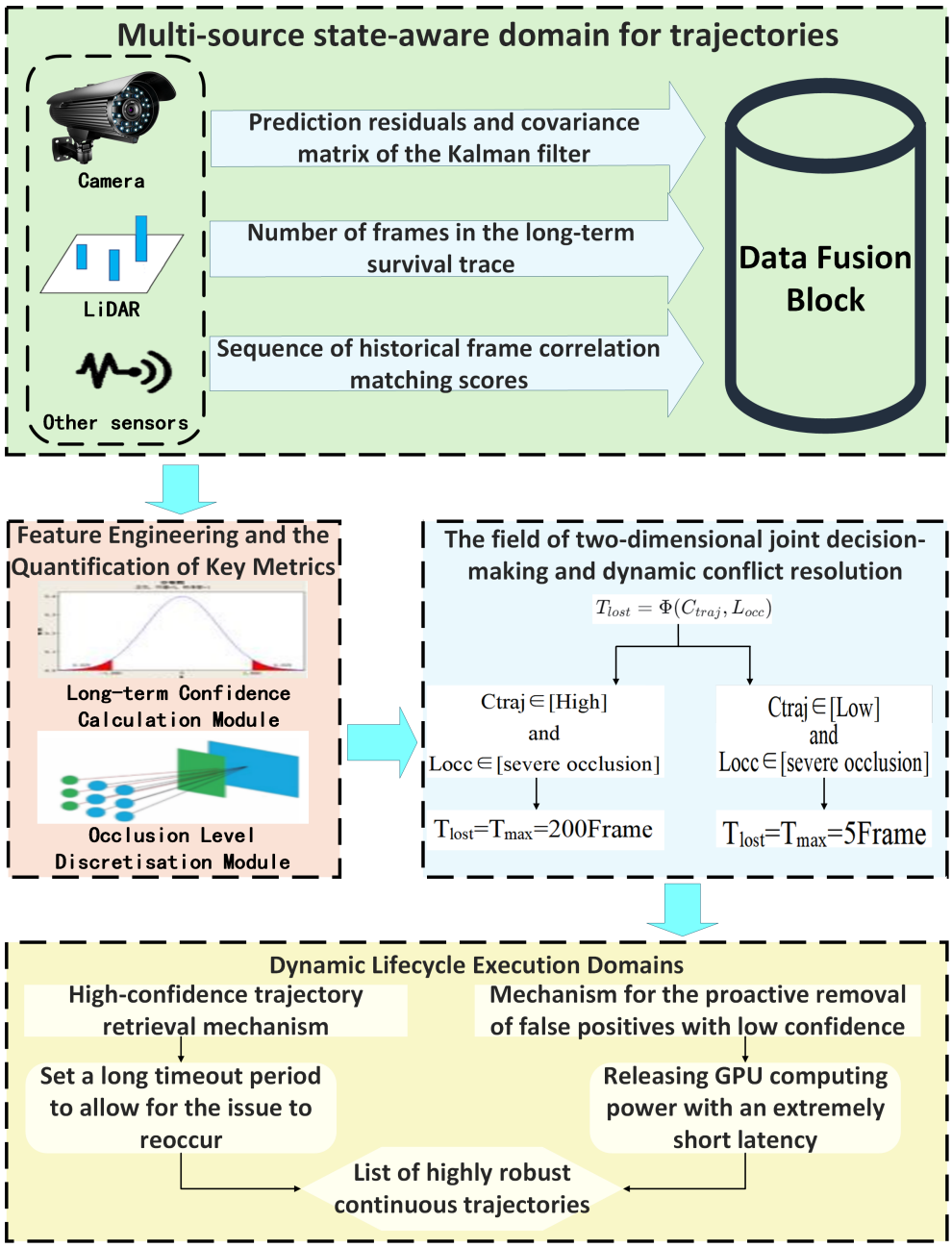


Fig. 5 Diagram of a four-tier dynamic trajectory management data flow architecture.

The system receives three types of data inputs in real time: the Kalman filter prediction residuals and covariance matrix, which characterise the uncertainty in the target's motion; the number of frames over which the trajectory has survived, T_{life}; and

the sequence of historical frame association matching scores, $\Sigma\text{Score}_{\text{hist}}$. These data streams are fed into the core "data fusion block"; after being normalised and aggregated, they provide the basis for the subsequent refined quantification of trajectory confidence and occlusion levels.

The second section defines the long-term trajectory confidence, C_{traj}. This confidence measure is a comprehensive reflection of both the historical match quality of the target and its survival time; its mathematical expression is as follows:

$$C_{\text{traj}} = \gamma \cdot \frac{1}{K} \sum_{t=1}^{K} \text{Score}_{\text{match},t} + (1-\gamma) \cdot \frac{T_{\text{life}}}{T_{\max}} \tag{8}$$

K represents the number of historical matching frames, T_{life} is the number of frames for which the trajectory has survived, $T_{\max}$ is the survival limit, γ is the balancing coefficient, and $\text{Score}_{\text{match},t}$ denotes the historical matching score for the tth frame.

Next, a 3×3 two-dimensional decision control matrix Φ is constructed using the occlusion severity L_{occ} of the trajectory (mapped by $S_{\text{occ}}(t)$ into low, medium and high occlusion levels) and the trajectory confidence C_{traj} (low, medium and high). The calculation rule for the differential loss tolerance time T_{lost} is as follows:

$$T_{\text{lost}} = \Phi(L_{\text{occ}}, C_{\text{traj}}) \tag{9}$$

Long-term confidence is denoted by C_{traj}, where $\Phi(\cdot)$ is the joint hierarchical decision mapping function. To achieve the objective of hierarchical decision-making with regard to the tolerance time for lost trajectories, this is represented by a two-dimensional joint decision matrix, as shown in Figure 6.

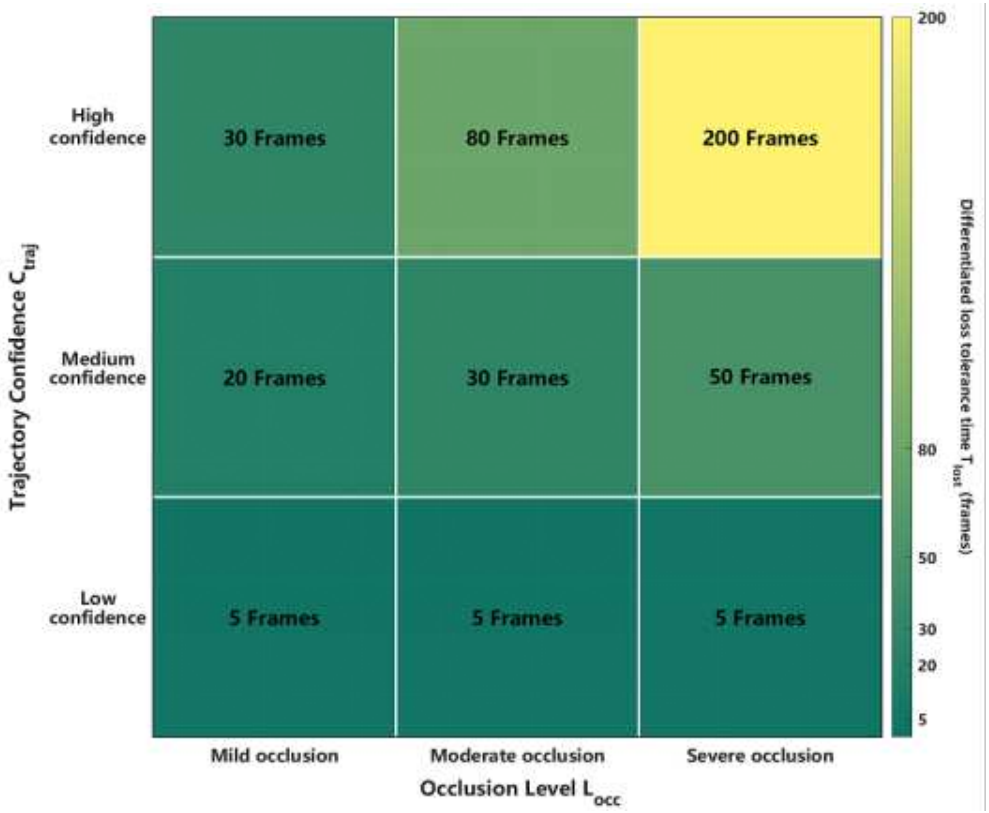


Fig. 6 Two-dimensional decision matrix for differentiated loss tolerance time T_{lost}(units: "frames")

Based on the weighting of the matrix, the system will implement the following tiered control measures: (1) Maximum protection (delayed termination): in cases of high confidence but severe occlusion, the system will allow the longest tolerance period (e.g. $T_{\text{lost}} = 200$ frames), resume the track only once the target has reappeared and has been successfully re-associated; (2) Medium protection (standard operation): where the confidence level is "medium" or "high" and the obstruction level is classified as minor or moderate, a corresponding tolerance period will be specified.; (3) Rapid Removal (Low-Confidence Filtering): For trajectories with low confidence, regardless of the occlusion situation, the system allocates a very short retention period (e.g. $T_{\text{lost}} = 5$ frames) before removing them from the database, thereby freeing up computational resources. The two-dimensional hierarchical management approach effectively overcomes the shortcomings of a globally uniform discard strategy, significantly improving the system's long-term tracking recall rate and real-time performance in environments where frequent high-level occlusions occur.

Finally, this differential strategy drives the underlying execution flow:

(1) A high-confidence trajectory retrieval mechanism: it tolerates long waiting periods for trajectories with 200 frames of output; when the target reappears within the

field of view, it immediately re-labels the original ID, thereby achieving trajectory re-association.

(2) The active clearance mechanism for low-confidence false positives immediately removes trajectories with an output of 5 frames from the cache, freeing up GPU video memory and computational resources.

The result is a continuous list of trajectories generated by a two-dimensional joint hierarchical framework characterised by high robustness, long-term stability, high recall and low system latency.

4 Experiments and Analysis

This section primarily employs experimental methods to verify the effectiveness of the proposed method in complex occlusion scenarios encountered in autonomous driving. We first describe the dataset and evaluation metrics used in the experiments, and then conduct a comprehensive evaluation of the proposed method's tracking accuracy, identity preservation capability and real-time performance through comparative experiments, module ablation studies and qualitative visualisation analyses.

4.1 Dataset

In order to conduct a comprehensive evaluation of multi-object tracking, two representative large-scale datasets—KITTI and nuScenes—were selected from the field of autonomous driving for analysis.

(1) **KITTI**, a classic 2D multi-object tracking benchmark in the field of autonomous driving, comprises 21 training sequences and 29 test sequences of real-world road images. The dataset encompasses a variety of traffic scenarios, including urban, rural and motorway environments, and provides detailed annotations for vehicles and pedestrians, thereby facilitating the quantitative evaluation of algorithms in typical urban and motorway settings.

(2) **nuScenes** is currently one of the largest publicly available datasets for autonomous driving, comprising 1,000 real-world driving scenes, officially categorised as 700 training scenes, 150 validation scenes and 150 test scenes. The dataset includes a variety of complex urban and rural traffic conditions, along with detailed annotations regarding visibility and the degree of occlusion, enabling quantitative evaluation of algorithms under conditions of severe occlusion.

In order to evaluate the proposed method more rigorously under "severe occlusion" conditions, the latest OccNuScenes benchmark [31] was utilised in the experiments. This benchmark provides a controlled, parameterised and reproducible occlusion generation protocol for three modalities—camera, radar and lidar—and also offers a dedicated occlusion dataset. The nuScenes subset experiments utilised the benchmark's "occlusion predefined labels" to group the tests, ensuring a high degree of specificity for "occlusion-aware" metrics.

4.2 Evaluation criteria

To ensure the objectivity and reproducibility of the experimental results, this paper adopts standard multi-object tracking and detection metrics, including: multi-object tracking accuracy (MOTA), multi-object tracking precision (MOTP—which is primarily described in terms of precision in the experimental tables), IDF1 score (IDF1), mean average precision (mAP), precision, recall and frame rate (FPS). In accordance with the metric selection criteria adopted in recent years for airborne perception and the evaluation of multi-agent dynamic trajectories [27][28], the rigorous mathematical definitions of each metric are as follows:

(1) **MOTA and IDF1:**

$$\mathrm{MOTA} = 1 - \frac{\sum_t (\mathrm{FN}_t + \mathrm{FP}_t + \mathrm{IDSW}_t)}{\sum_t \mathrm{GT}_t} \tag{10}$$

$$\mathrm{IDF1} = \frac{2 \times \mathrm{IDTP}}{2 \times \mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}} \tag{11}$$

GT_t denotes the actual number of targets in frame t, whilst IDTP, IDFP and IDFN denote the total numbers of correctly identified, incorrectly identified and missed targets, respectively, based on identity recognition matching.

(2) **mAP, Precision and Recall:**

Let TP denote the number of correct detections, FP the number of false positives, and FN the number of false negatives, the expressions for Precision and Recall are defined as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{12}$$

The average precision, expressed as the area under the precision-recall curve, is given by the following formula:

$$\mathrm{AP} = \sum_{n=1}^{N} (\mathrm{R}_n - \mathrm{R}_{n-1}) \cdot \mathrm{P}_n \tag{13}$$

Here, P_n is the precision at the nth recall threshold, and R_n is the recall at the nth recall threshold. mAP is the average of the AP values across all classes (e.g. vehicles, pedestrians): $\mathrm{mAP} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{AP}_c$.

(3) **FPS frame rate:**

FPS is used to assess the feasibility of deploying an algorithm in real time on in-vehicle devices. It is defined as the reciprocal of the average inference time required to process a single frame of imagery:

$$\mathrm{FPS} = \frac{1}{t_{\mathrm{inf}}} \tag{14}$$

The average time taken for a single frame of an image to undergo feature extraction, neural network inference and processing is t_{inf}.

All experiments in this paper were implemented using the PyTorch framework, with training and inference carried out on a single NVIDIA RTX 4090 GPU. YOLOv8x was used as the detection backbone network, with a fixed input image resolution of 1280×720. With regard to object detection in occluded scenes on the KITTI dataset, the research by Yusuf [39] et al. utilised attention modules (CBAM/CoordAtt) to improve the detection accuracy of YOLOv8n in KITTI occlusion scenarios. Whilst this paper primarily focuses on the tracking and management of detectors, in order to avoid errors inherent to the detectors themselves, all compared algorithms utilise YOLOv8x with identical configurations as a common upstream input.

4.3 Comparative Analysis

To demonstrate the advanced nature of the method proposed in this paper, a comparison was carried out with five classical and recent multi-object tracking algorithms mentioned in the paper, namely SORA-SORT, CLIP-DeepSORT, LiDAR-ByteTrack, Dual-Threshold OC-SORT and EKF-BoT-SORT. The specific improvement strategies of these five comparative methods are as follows: SORA-SORT, proposed by Kato[33] et al. employs a simple occlusion risk-aware framework to enhance the robustness of multi-object tracking; CLIP-DeepSORT, proposed by Alkandary[34] et al. incorporates the CLIP vision-language large model into the re-identification modules of DeepSORT and StrongSORT, thereby strengthening the ability to distinguish appearance features in complex scenes; LiDAR-ByteTrack, proposed by Mohibullah[35] et al., utilises ankle-level 2D LiDAR data in conjunction with the ByteTrack framework to improve pedestrian tracking in crowded scenes; Dual-Threshold OC-SORT, proposed by Zhao[36] et al., employs a dual-threshold correlation approach to enhance real-time tracking performance from a drone's perspective whilst ensuring a lightweight system; Arief[37] et al. proposed EKF-BoT-SORT, which utilises an extended Kalman filter to handle non-linear motion within uncertain mixed traffic flows, thereby enhancing the tracking stability of BoT-SORT.

The detection inputs for all comparison algorithms were uniformly based on the detection results from YOLOv8, and testing was carried out on the same hardware environment. The quantitative comparison results for the KITTI and nuScenes datasets are shown in Tables I and II respectively.

Table 1 Comparison of tracking and detection performance across different algorithms on the KITTI dataset

Methods	**MOTA↑**	**IDF1↑**	**mAP↑**	**Precision↑**	**Recall↑**	**FPS↑**
SORA-SORT[33]	48.3	51.7	62.1	71.4	68.2	165.2
CLIP-DeepSORT[34]	56.2	58.4	65.8	76.5	72.3	56.7
LiDAR-ByteTrack[35]	63.5	65.3	67.2	78.9	75.6	108.9
Dual-Threshold OC-SORT[36]	64.1	66.7	66.5	78.3	76.1	72.4
EKF-BoT-SORT[37]	66.4	68.5	67.8	80.2	77.5	23.8
Methodology of this paper	**69.8**	**72.2**	**68.5**	**81.6**	**79.3**	**61.4**

Table 2 Comparison of tracking and detection performance across different algorithms on the nuScenes dataset

Methods	MOTA↑	IDF1↑	mAP↑	Precision↑	Recall↑	FPS↑
SORA-SORT[33]	42.5	44.1	48.3	57.9	53.7	152.5
CLIP-DeepSORT[34]	49.3	51.8	53.7	62.1	59.4	49.8
LiDAR-ByteTrack[35]	58.2	59.7	55.4	65.5	61.2	98.2
Dual-Threshold OC-SORT[36]	57.9	60.4	54.9	64.7	60.8	65.3
EKF-BoT-SORT[37]	60.5	62.8	56.7	66.9	63.7	19.4
Methodology of this paper	**64.3**	**66.9**	**57.5**	**68.4**	**66.1**	**54.1**

As can be seen from Tables 1 and 2, the method proposed in this paper achieved leading MOTA and IDF1 scores on both the KITTI and nuScenes datasets.On the KITTI dataset, our method achieved a MOTA of 69.8 per cent, representing an improvement of 3.4 per cent over EKF-BoT-SORT[37], the best-performing method among the baselines; on the nuScenes dataset, the MOTA reached 64.3 per cent, representing an improvement of 3.8 per cent over EKF-BoT-SORT[37], the best-performing method among the baselines.

It can be seen from the above results that although these improved variants are tailored to specific scenarios such as drones, LiDAR and mixed traffic flows, they still lack a dynamic, adaptive perception method for situations where targets are occluded . Although SORA-SORT[33] takes into account the detection of occlusion risks, it still employs fixed weights or global thresholds for association and updating, making it difficult to effectively suppress template contamination in extremely dense traffic flows;EKF-BoT-SORT[37] improves upon the motion model; however, in cases of prolonged severe occlusion, the lack of a mechanism for differentiated and precise management of the trajectory lifecycle can easily lead to premature trajectory breakage . The method described in this paper utilises techniques such as continuous occlusion score-driven multi-feature dynamic weighting, forced template update freezing, and joint hierarchical trajectory management to overcome the shortcomings mentioned earlier. Whilst maintaining real-time inference speeds of 61.4 FPS (KITTI) and 54.1 FPS (nuScenes), it also achieves significant improvements in tracking accuracy and identity preservation.

To clearly demonstrate the advantages of the proposed method for tracking in occlusion scenarios, a visual comparison was carried out based on a selection of typical occlusion scenarios. The figure shows the tracking results of the two methods across three consecutive frames. In the baseline method on the left, some target IDs have changed between frames, and the trajectories of a few targets have been interrupted, only to reappear with new IDs. In contrast, in the method proposed in this paper on the right, the target IDs remain constant throughout the process, and both the detection boxes and motion trajectories remain consistently stable. It is thus immediately apparent that the method proposed in this paper offers a clear advantage in tracking scenarios involving dense multi-target interactions. Tracking in a multi-object motorway scenario is shown in Figure 7 below.

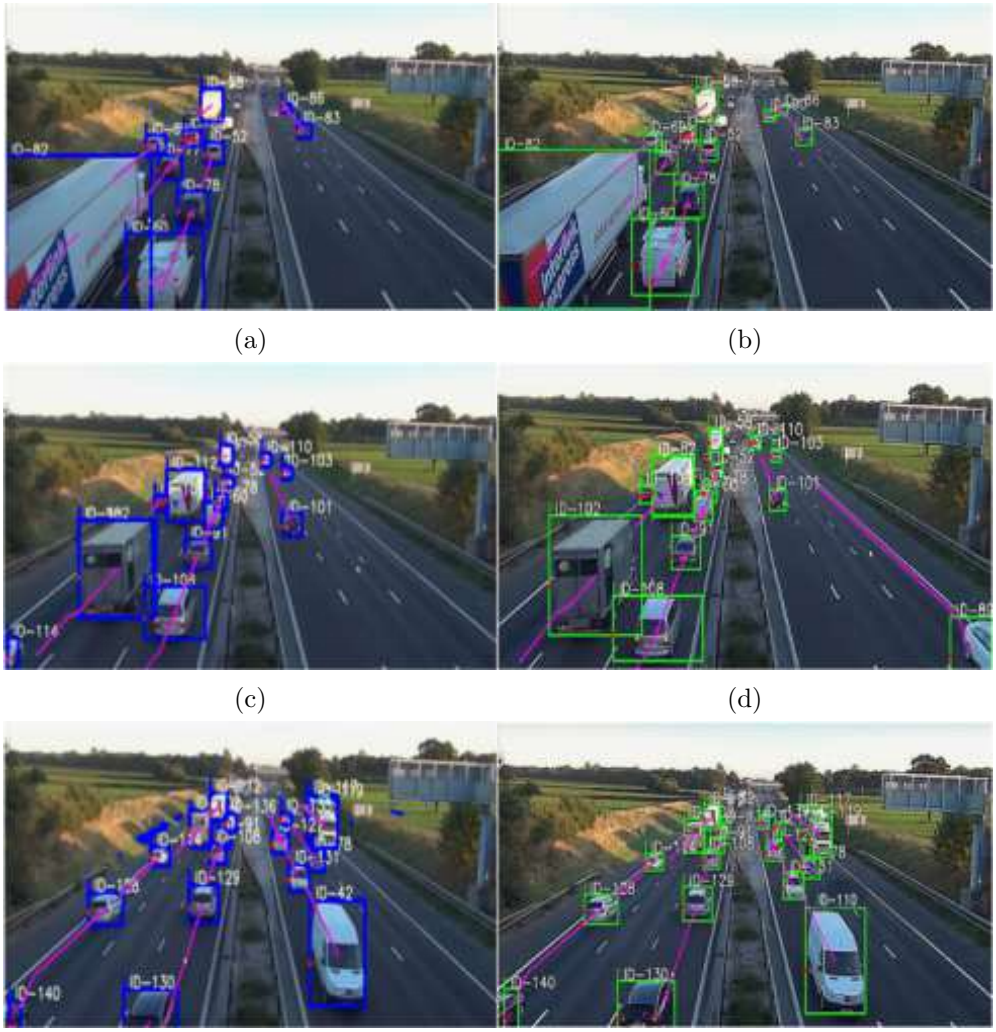


Fig. 7 Comparison of multi-target scene tracking on motorways.

The differences shown in Figure7 stem from the fundamental differences between the two methods in terms of data association and trajectory management mechanisms. In motorway scenarios, where there are a large number of targets travelling at high speeds, targets frequently come close to one another, are briefly occluded, or appear similar in appearance; consequently, high robustness is required for data association. The baseline method uses fixed weights to fuse appearance, motion and geometric features during the association stage, and applies the same loss tolerance threshold to all trajectories: that is, when two targets come into close proximity for a brief period, causing their detection boxes to overlap, the fixed weights make it easy for matches between visually similar targets to go wrong; a single incorrect match can

then propagate along the trajectory chain, leading to a chain reaction of ID swaps; Furthermore, when a target temporarily disappears due to occlusion or a missed detection, the uniform tolerance threshold causes the trajectory to terminate as soon as it is reached; when the target reappears, it must be reinitialised with a completely new ID, resulting in trajectory breaks. The method proposed in this paper considers three aspects—association, templates and trajectories—in unison. Specifically, it dynamically adjusts feature weights based on the continuous occlusion score; when a target approaches or is occluded, it automatically reduces the weight of unreliable appearance metrics whilst increasing the influence of motion and geometric IoU, thereby preventing false matches at source; The ODTU strategy freezes updates to the appearance template when the occlusion score exceeds the threshold ($S_{\text{occ}} \geq 0.7$), preventing the appearance features of similar targets from interfering with identity information; the OLC-HTManage framework makes decisions based on a combination of occlusion severity and trajectory confidence, granting longer loss tolerance times to high-confidence trajectories, allowing targets to retain their original IDs and continue their trajectories even after brief disappearances. The combined effect of these three approaches ensures that the IDs of each target on the right-hand side of Figure 7 remain stable and their trajectories remain continuous and complete, whereas the baseline method exhibits ID jumps and trajectory breaks.

We validate the robustness of the proposed method in scenarios with dense pedestrian occlusion by conducting a visual comparison using scenes at busy road junctions. Figure 8 illustrates the tracking results of the two methods across three consecutive frames: in the baseline method on the left, when pedestrians occlude one another, some object bounding boxes become mismatched with their ID annotations, and after a break in the trajectory, the pedestrian reappears with a new identity; on the right, in our method, the IDs of each pedestrian target remain consistent before and after occlusion, and the detection bounding boxes and trajectory lines remain continuous and stable, intuitively demonstrating the advantage of our method in maintaining identity under dense pedestrian occlusion scenarios. A comparison of tracking in dense pedestrian occlusion scenarios is shown in Figure 8 below.



Fig. 8 Comparison of tracking in scenes with dense pedestrian occlusions.

As can be seen from Figure8, as the pedestrian's movement and occlusion intensify, the shortcomings of the Baseline method gradually worsen: in Figure (a), detection boxes become fragmented and false background detections occur; in Figure (c), ID drift and box displacement occur; and in Figure (e), fragmentation and false alarms become significantly more severe, whilst ID numbering becomes chaotic; In contrast, the method proposed in this paper maintains continuous and consistent core target IDs and intact bounding boxes across all three frames (Figures b, d and f), with no instances of ID switching. This difference stems from the distinct data association and trajectory management mechanisms employed by the two methods. The baseline employs only fixed-weight association, a fixed update rate and a uniform loss

tolerance time. Dense occlusions cause appearance features to be continuously contaminated; fixed weights cannot distinguish between reliable and unreliable features, whilst a uniform threshold causes occluded trajectories to terminate prematurely, thereby generating a large number of trajectory breaks and identity jumps. In contrast, the method proposed in this paper progressively addresses these issues through the coordinated design of occlusion-aware adaptive association, occlusion-driven template updating, and joint hierarchical trajectory management. The occlusion-aware adaptive association mechanism automatically reduces the weight of appearance features during occlusion whilst strengthening the fusion weight of motion and IoU, thereby reducing false matches; the occlusion-driven template updating strategy freezes templates during severe occlusion, preventing cross-contamination of pedestrian appearance features; the joint hierarchical trajectory management grants longer survival times to high-confidence occluded trajectories, enabling correctly re-identified pedestrians to be retrieved after re-appearance. Through the synergy of these three components, the method significantly reduces ID switching in densely populated pedestrian scenes whilst maintaining trajectory continuity and identity consistency.

To evaluate the performance of the proposed method in re-identifying targets that briefly leave and then re-enter the field of view, Figure 9 shows the tracking results for both methods before and after occlusion, as well as after re-entry. In the baseline method on the left, ObjectID1 was not detected whilst within the occlusion region; upon re-entry, it was incorrectly assigned a new identity (ObjectID4), resulting in the loss of its trajectory. On the right, the proposed method is able to re-establish the original ID after the target re-enters the field of view, thereby achieving re-identification and ensuring trajectory continuity. A comparison of tracking performance in scenarios where targets enter and exit the field of view, including occlusion, is shown in Figure 9.

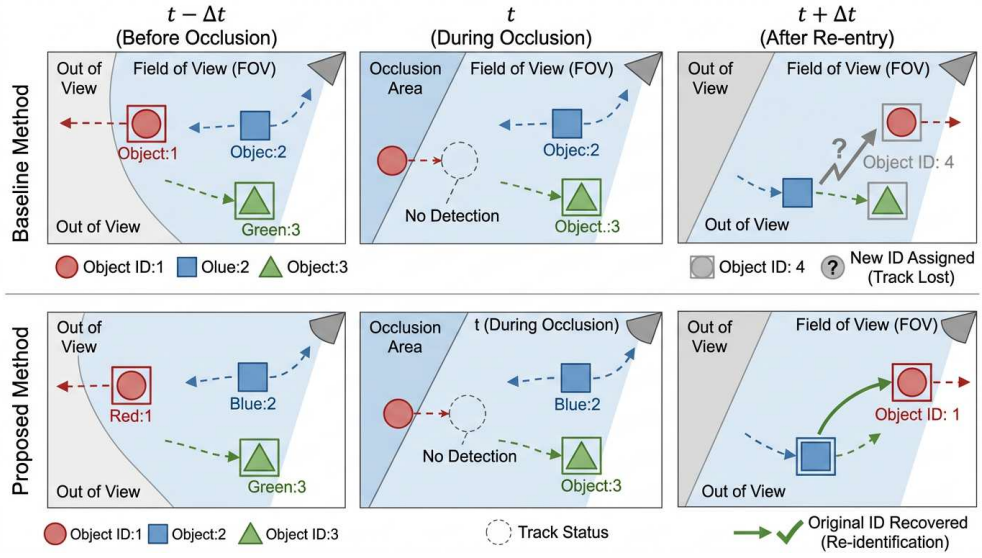


Fig. 9 Comparison of Target Entry and Exit, Field-of-View Occlusion and Scene Tracking.

Figure9 illustrates the differences in the re-identification mechanisms employed by various methods following a prolonged loss of the target. The baseline method uses a uniform loss tolerance threshold to determine that ObjectID1 has not been detected for a continuous period because it has entered an occlusion zone; consequently, the trajectory is terminated within a short period of time. When the target reappears, as the historical template has either been deleted along with the trajectory or was partially corrupted prior to the occlusion, appearance matching cannot be performed. The tracker is therefore forced to treat the target as a new object and assign it a new ObjectID4, resulting in a trajectory break and an identity switch.

The method described in this paper effectively resolves this issue by combining a joint hierarchical trajectory management framework with an occlusion-driven template update strategy. On the one hand, joint hierarchical trajectory management is employed to make two-dimensional decisions based on occlusion levels and trajectory confidence; high-confidence occluded trajectories are assigned a very long loss tolerance time ($T_{\text{lost}} = 200$ frames) to ensure that the trajectory of ObjectID1 remains active and is not terminated prematurely; on the other hand, the occlusion-driven appearance template update strategy forces the template to be frozen during occlusion, preserving the object's pre-occlusion appearance features, thereby ensuring that appearance similarity matching can still be reliably performed when the object re-enters the frame. By combining these two approaches, the proposed method links the newly detected and surviving trajectories to ObjectID1 at time t+At, restoring the original identity and continuing the trajectory, thereby preventing the generation of a new ID. This result is consistent with the quantitative conclusion of a significant reduction in IDSwitch shown in Table 2, further demonstrating the effectiveness of the method described in this paper when the target enters and exits the field of view and is occluded.

4.4 Ablation experiments

Through ablation experiments involving the removal of individual modules on the KITTI dataset, we have further validated the independent contributions of the three core innovations presented in this paper. The baseline algorithm (Baseline) is the standard ByteTrack framework, which utilises only fixed weights, a fixed update rate and a uniform loss tolerance time. The results of the ablation experiments on the KITTI dataset are shown in TableIII below.

Table 3 Results of ablation experiments on the KITTI dataset.

Algorithm Configuration	MOTA ↑	IDF1 ↑	mAP ↑	Prec ↑	Recall ↑	FPS ↑	Change (ΔMOTA)
Baseline	63.1	65.0	67.2	78.4	75.2	108.9	-
+ Occlusion-Aware Adaptive Correlation	65.8	67.9	67.9	80.0	77.1	82.4	+2.7
+ Update to the Occlusion Driver Template	66.2	69.1	68.1	80.3	78.2	75.6	+3.1
Methodology of this paper	69.8	72.2	68.5	81.6	79.3	61.4	+6.7

Ablation Analysis:

(1) When incorporating occlusion-aware adaptive association, MOTA improved by 2.7 per cent and mAP by 0.7 per cent. This indicates that, in cases of severe occlusion, this mechanism reduces the weighting of appearance whilst increasing the combined weighting of motion and IoU, thereby reducing false matches and ID switches caused by appearance-based misclassification and lowering the number of false positives, which in turn leads to an initial improvement in mAP.
(2) When incorporating occlusion-driven template updates, MOTA improved by 3.1 per cent, IDF1 by 4.1 per cent, and mAP by 0.9 per cent. Thanks to the 'freeze-update' strategy, background noise injection into feature templates was effectively suppressed, significantly reducing ID switches following prolonged loss. The slight improvement in mAP is attributable to the fact that when the target reappears, the system can quickly re-identify it because the template remains uncontaminated, thereby significantly reducing false negatives (FN) caused by ID mismatches.
(3) Following the incorporation of joint hierarchical trajectory management, the model reached the final stage of the method described in this paper, achieving a MOTA of 69.8 per cent and an mAP of 68.5 per cent. The occlusion-aware score, Socc formed a robust 'data-driven closed loop' throughout the entire process, ensuring the algorithm's optimal performance in complex occlusion scenarios.

5 Conclusions

This paper addresses the issue of cascading failures in multi-object tracking for autonomous driving under complex occlusion conditions by proposing an occlusion-aware adaptive multi-feature fusion tracking framework. This framework constructs an evaluation space based on three dimensions: "occlusion-aware association → occlusion-driven updates → joint hierarchical management". It comprises three main innovative modules: an occlusion-aware adaptive multi-feature association mechanism, an occlusion-driven appearance template update strategy, and a joint hierarchical trajectory management framework. Experimental results from the large-scale KITTI and nuScenes datasets demonstrate that this algorithm exhibits significant advantages over existing mainstream methods in terms of metrics such as MOTA and IDF1, whilst effectively reducing ID switching. The complete framework presented in this paper achieves a significantly higher recall rate under severe occlusion than a single baseline strategy. Despite incorporating multiple modules for dynamic computation, the inference speed remains at 54-61 FPS on the two datasets, thereby meeting the dual stringent requirements of high precision and real-time performance demanded by autonomous driving systems.

Current methods still suffer from a certain degree of performance degradation in the face of extremely adverse weather conditions, prolonged occlusions lasting several seconds or more, and the absence of motion prior knowledge.Future work will focus on the vertical development and horizontal expansion of the current framework; specifically, this involves investigating the integration of multimodal information such as LiDAR, extending the multi-feature correlation described in this paper from a two-dimensional visual space to a three-dimensional geometric space, and enhancing robustness under extreme conditions; We will investigate self-supervised occlusion scoring methods to

simplify the system design of the current occlusion detection module and improve deployment efficiency; over a longer time horizon, we will further explore trajectory-level semantic information, extending the current hierarchical trajectory management to global semantic correlation and long-term trajectory prediction, whilst investigating its application in AI-based mobility management systems[29], thereby constructing a more comprehensive perception system for autonomous driving.

Acknowledgements. This work was supported in part by the 2025 Guangxi University Young and Middle-Aged Teachers basic Scientific Research Ability improvement Project "Research on multi-Object Tracking Fusion Model of Autonomous Vehicle"under Grant 2025KY0991.

Declarations

- Competing Interests: The authors have no relevant financial or non-financial interests to disclose.
- Author contribution: Yongjie Zhang: Supervision, Funding acquisition, Project administration, Conceptualization. Yanfeng Wu: Writing - original draft, Methodology, Visualization, Resources. Zi Gao: Writing - original draft, Visualization, Resources, Data curation, Software. Zeyue Li: Project administration, Writing - review & editing. Yumei Zhou: Project administration, Writing - review & editing. All authors have read and approved the final manuscript.

References

[1] Kumar S, Sharma S, Eising C (2025) Exploring sensor impact and architectural robustness in adverse weather on BEV perception. *IEEE Open Journal of Vehicular Technology* 6: 2857–2875. https://doi.org/10.1109/OJVT.2025.3621862.

[2] Lei Y, Shan C, Ge WC (2025) EdgeCaseDNet: An enhanced detection architecture for edge case perception in autonomous driving. *PLOS ONE* 20(12): e0338638. https://doi.org/10.1371/journal.pone.0338638.

[3] Zhang P, Wang D, Lu H (2024) Multi-modal visual tracking: Review and experimental comparison. *Journal of Ambient Intelligence and Humanized Computing* 10(2): 193–214. https://doi.org/10.1007/s41095-023-0345-5.

[4] Liu Z, Huang X, Sun J, et al. (2024) AMtrack: Anti-occlusion multi-object tracking algorithm. *Signal, Image and Video Processing* 18(12): 9305–9318. https://doi.org/10.1007/s11760-024-03547-w.

[5] Ogunrinde I, Bernadin S (2024) Improved DeepSORT-Based Object Tracking in Foggy Weather for AVs Using Sematic Labels and Fused Appearance Feature Network. *Sensors*. https://doi.org/10.3390/s24144692.

[6] Zhuang ZH, Yin MF, Guo XT, et al. (2026) Adaptive multi-level background perception and decision fusion RGBT object tracking. *Computer Engineering and Applications*. https://doi.org/10.3778/j.issn.1002-8331.2503-0167.

[7] Zhai H, Yan HY, Zhou JY, et al. (2025) InteBOMB: Integrating generic object tracking and segmentation with pose estimation for animal behavior analysis. *Zoological Research* 46(2): 355–369. https://doi.org/10.24272/j.issn.2095-8137.2024.268.

[8] Chen SF, Chen JC, Jhuo IH, Lin YY (2026) GOT-JEPA: Generic object tracking with model adaptation and occlusion handling using joint-embedding predictive architecture. *IEEE Transactions on Circuits and Systems for Video Technology* 36(7): 10836–10851. https://doi.org/10.1109/TCSVT.2026.3675005.

[9] Jin J, Zhang J, Zhang K, et al. (2024) 3D multi-object tracking with boosting data association and improved trajectory management mechanism. *Signal Processing* 218. https://doi.org/10.1016/j.sigpro.2023.109367.

[10] Wang H, Chen T, Wang Y (2025) Towards Occlusion-Aware Multi-Pedestrian Tracking. *Applied Sciences* 15(24). https://doi.org/10.3390/app152413045.

[11] Krejčí J, Kost O, Xia Y, et al. (2025) Occlusion-aware multi-object tracking via expected probability of detection. *IEEE Transactions on Aerospace and Electronic Systems*. https://doi.org/10.1109/TAES.2026.3676431.

[12] Tsai CY, Shen GY, Nisar H (2023) Swin-JDE: Joint Detection and Embedding Multi-Object Tracking in Crowded Scenes Based on Swin-Transformer. *Engineering Applications of Artificial Intelligence* 119. https://doi.org/10.1016/j.engappai.2022.105770.

[13] Chen Y, Tang Y, Xiao Y, et al. (2024) Satellite video single object tracking: A systematic review and an oriented object tracking benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing* 210: 212–240. https://doi.org/10.1016/j.isprsjprs.2024.03.013.

[14] Huang K, Hao Q (2021) Joint Multi-Object Detection and Tracking with Camera-LiDAR Fusion for Autonomous Driving. *arXiv*. https://doi.org/10.48550/arXiv.2108.04602.

[15] Tan S, Li X, Xu Q, et al. (2025) 3D multi-object tracking based on parallel multimodal data association. *Machine Vision and Applications*. https://doi.org/10.1007/s00138-025-01675-y.

[16] Sun S, Wang C, Xiao B, et al. (2024) Enhanced Kalman Filter with Dummy Nodes and Prediction Confidence for Bipartite Graph Matching in 3D Multi-Object Tracking. *Electronics* 13(24). https://doi.org/10.3390/electronics13244950.

[17] Liu L, Wang JX, Li CH, Li CL (2026) Unveiling the power of multi-modal template update in RGBT tracking. *IEEE Transactions on Image Processing* 35: 6026–6040. https://doi.org/10.1109/TIP.2026.3698355.

[18] Sakthi KS, Jeong JH, Choi WY (2026) DynaTrack: Dynamic memory-enhanced transformer with deformable attention for robust single object tracking. *IEEE Access* 14: 87669–87684. https://doi.org/10.1109/ACCESS.2026.3700286.

[19] Li R, Zhang D, Wang Y, et al. (2025) Open-vocabulary multi-object tracking with domain generalized and temporally adaptive features. *IEEE Transactions on Multimedia* 27: 3009–3022. https://doi.org/10.1109/TMM.2025.3557619.

[20] Zhu J, Chen X, Zhang PY, et al. (2024) SRRT: Exploring search region regulation for visual object tracking. *IEEE Transactions on Circuits and Systems for Video Technology* 34(11): 10551–10563. https://doi.org/10.1109/TCSVT.2024.3409898.

[21] Gamerdinger J, Teufel S, Volk G, et al. (2023) Analyzing track management strategies for multi object tracking in cooperative autonomous driving scenarios. *at - Automatisierungstechnik* 71(4): 287–294. https://doi.org/10.1515/auto-2022-0157.

[22] Min Z, Hassan GM, Jo GS (2024) Rethinking motion estimation: an outlier removal strategy in SORT for multi-object tracking with camera moving. *IEEE Access* 12: 142819–142837. https://doi.org/10.1109/ACCESS.2024.3432156.

[23] Zeng L, Chen H, Liu Y, et al. (2026) SORT-LFR: Revisiting SORT for multi-object tracking in low-frame-rate videos. *IEEE Transactions on Multimedia* 28: 3364–3379. https://doi.org/10.1109/TMM.2026.3651130.

[24] Li Y, Jiang Y, Xiong Z, Wu X (2024) Improving interaction-based vehicle trajectory prediction via handling sensing failures. *IEEE Sensors Journal* 24(14): 22907–22915. https://doi.org/10.1109/JSEN.2024.3405542.

[25] Zhang ZM, Ye C (2026) Drone multi-object tracking algorithm based on improved ByteTrack and YOLOv10. *Acta Armamentarii* 47(1): 261–271. https://doi.org/10.12382/bgxb.2025.0609.

[26] Luo B, Sun J, Tang R, et al. (2025) Reinforcement learning-based 3D trajectory tracking control of hypersonic gliding vehicles with time-varying uncertainties. *IEEE Transactions on Automation Science and Engineering* 22: 8187–8199. https://doi.org/10.1109/TASE.2024.3481422.

[27] Wang Z, Wei T, Sun G, et al. (2025) Multi-UAV Enabled MEC Networks: Optimizing Delay Through Intelligent 3-D Trajectory Planning and Resource Allocation. *IEEE Transactions on Intelligent Transportation Systems* 26(11): 20897–20911. https://doi.org/10.1109/TITS.2025.3586874.

[28] Vrba M, Walter V, Pritzl V, et al. (2025) On onboard LiDAR-based flying object detection. *IEEE Transactions on Robotics* 41: 593–611. https://doi.org/10.1109/TRO.2024.3502494.

[29] Jahanmanesh P, Farhadi A, Zamanifar A (2025) Mobility management with AI. *IEEE Access* 13: 83022–83060. https://doi.org/10.1109/ACCESS.2025.3566337.

[30] Hao Y, Chung H, Vu HL (2025) Integrating local planners into autonomous intersection management: Bridging global planning and dynamic execution. *IEEE Transactions on Intelligent Vehicles* 10(10): 4642–4654. https://doi.org/10.1109/TIV.2024.3488529.

[31] Kumar S, et al. (2026) Descriptor: Occluded nuScenes Dataset (OccNuScenes). *IEEE Data Descriptions* 3: 149–158. https://doi.org/10.1109/IEEEDATA.2026.3667122.

[32] Zhang Y, Sun P, Jiang Y, et al. (2022) ByteTrack: multi-object tracking by associating every detection box. In: *Proceedings of the European Conference on Computer Vision*. Springer, Cham, pp 1–21. https://doi.org/10.1007/978-3-031-20047-2_1.

[33] Kato K, Goto D, Huang WH, et al. (2025) SORA-SORT: A simple occlusion risk-aware framework for multi-object tracking. In: *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, New York, pp 251–255. https://doi.org/10.1145/3714394.3754385.

[34] Alkandary K, Yildiz AS, Meng HY (2026) Enhancing multi object tracking with CLIP: a comparative study on DeepSORT and StrongSORT. *Electronics* 15(2): 265. https://doi.org/10.3390/electronics15020265.

[35] Mohibullah M, Hironaka Y, Suda Y, et al. (2025) Pedestrian tracking using ankle-level 2D-LiDAR based on ByteTrack. In: *Advances in Visual Computing, ISVC 2024, Pt I*. Springer, Cham, pp 211–222. https://doi.org/10.1007/978-3-031-77392-1_16.

[36] Zhao Y, Wang X, Zhou X, et al. (2025) Dual-threshold association strategy for enhancing UAV-based object tracking with OC-SORT: A lightweight approach for real-time UAV multi-object tracking. In: *2025 3rd International Conference on Signal Processing and Intelligent Computing (SPIC)*. IEEE, Piscataway, pp 1–6. https://doi.org/10.1109/SPIC68204.2025.11470761.

[37] Arief M, Afdhal A, Nasaruddin N (2025) Real-time detection and tracking framework using extended Kalman filter BoT-SORT in uncertainty mixed-traffic. *IEEE Open Journal of Vehicular Technology* 6: 2804–2827. https://doi.org/10.1109/OJVT.2025.3617470.

[38] Liu XZ, Liu H, Wang KL, et al. (2025) StrongerCenter: Enhancing TransCenter for robust multi-object tracking. *Neurocomputing* 657: 131527. https://doi.org/10.1016/j.neucom.2025.131527.

[39] Yusuf MM, Nur Ramadhani K, Firdaus F (2025) CBAM and CoordAtt in YOLOv8n: Attention Location Effects on KITTI Occlusion Detection. In: *2025 International Conference on Information Technology and Computing (ICITCOM)*. IEEE, Yogyakarta, pp 89–93. https://doi.org/10.1109/ICITCOM66635.2025.11265080.